DOCUMENTATION
INCORPORATED

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

THE PREPARATION OF MANUAL DICTIONARIES OF ASSOCIATION

TECHNICAL REPORT NO. 5

PREPARED UNDER

CONTRACT NO. Nonr-1305(00)

for

THE OFFICE OF NAVAL RESEARCH

APRIL
1954

TECHNICAL REPORT NO. 5

THE PREPARATION OF MANUAL DICTIONARIES OF ASSOCIATION

In the previous report, we explained the logical significance of the symbol "*" to be read "is associated with" and the symbol "," to be read as "and" between classes. In this paper, we will describe the method of preparing manual dictionaries of association for any system of information.

Let us consider a system of information, S, comprising N items of information (documents, reports, books, etc.) indexed by the terms $t_1, t_2, \ldots, t_n$. The sets of terms used to analyze or index the items of information will be designated by $I_1, I_2, \ldots, I_N$, the subscript of any I indicating the number of the document indexed by that set of terms. The set of terms associated with any given term t_r we will call P_r: $t_k \in P_r \equiv t_k * t_r$. Thus P_1, P_2, etc. will constitute the "pages" of our dictionary. We shall write $t_k * P_r$ if t_k is associated with each term of P_r.

The three elements which determine the size of any system are:

The number of items, N

The number of terms, n

The average number of terms used to analyze each item, $\frac{\sum_k (\text{number of terms in } I_k)}{N}$

"$\in$" means "belongs to"

As in <u>Technical Report No. 3</u>, we shall assume that S contains 50,000 items analyzed by $I_1, I_2, \ldots, I_{50,000}$; and that each I is a unique set of 10 terms chosen from $t_1, t_2, \ldots, t_{5,000}$.

There are many ways to organize an S so that the actual associations of the system can be exhibited. But it will be recalled that the number of actual associations in a system of 50,000 items each analyzed by 10 terms will range from 25,000,000 to 50.000.000. This means that although we could arrange all the associations in a linear alphabetical card file, the file would contain upwards of 25,000,000 cards and would be too big to be useful. We could also set up a file in which each I was recorded just once. Such a file would contain only 50,000 cards. But we would not know how to arrange it, since each I has 1023 potential filing positions. As a random IBM file, the search for any t_r would involve the sorting of 50,000 items. Such a search would yield a group of I's containing t_r.

Let us assume that such a search yielded 200 I's. The number of terms in such a group would be 2000 but many of these might be duplicates. We would thus have to examine the 200 cards to list the different t's and eliminate duplicates. If now we asked for all t_k's for which $t_k * t_r * t_s$, we could resort our 200 I's to determine which ones contain both t_r and t_s. This whole process is also too laborious to be a practical method.

The method we have chosen for the initial preparation of dictionaries which exhibit the association of any group is as follows:

Suppose any term t_r appears in sets $I_1, I_2, \ldots, I_{200}$, we then construct P_r, which includes all distinct terms contained in any set which contains t_r, that is, all the terms of $I_1, I_2, \ldots, I_{200}$. Using materials we had indexed for another project, we selected three terms from the vocabulary set, namely, Waves (t_a); Boundary (t_b); and Lamination, Laminar (t_c).

For each of these terms we constructed a "page" by listing in sequence all other terms associated with them. t_a appeared on 52 items and in association with 247 other terms. This indicates that many terms were duplicated in analyzing the 52 items. Otherwise the P_a associated with t_a would have contained 468 terms. t_b appeared on 43 items, associated with 204 different terms; and t_c appeared on 40 items associated with 160 terms. The three "pages" are presented as Exhibits 1, 2, and 3.

If we enumerate the members of P_a, we have $t_a * (t_1, t_2, t_3, \ldots)$. Since $A * A$, $t_a \in P_a$. Further, if $t_a * t_b$, then both P_a and P_b will contain t_a and t_b. Thus, if "waves" is associated with "laminar" in S, we would discover on the page for "waves" and on the page for "laminar" the terms "waves" and "laminar".

Suppose we write out (Exhibit 4) all the terms common to P_a and P_b. The terms common to P_a and P_b constitute the logical product of the two sets and we can always write $(t_a, t_b) * (P_a X P_b)$, but not $(t_a * t_b) * (P_a X P_b)$. That is, the terms which are common to the two lists are associated with "waves" and with "laminar", but not necessarily with both at once (in the same document).

Suppose now we write out the words which are common to our three pages (Exhibit 5), that is $(t_a, t_b, t_c) *$ $[(P_a X P_b) X P_c]$. Our result will be

$$(t_a, t_b, t_c) * (t_a, t_b, t_c, \ldots)$$

since each is associated with the other two. Here again, it is important to note that we still have not advanced beyond a chain of pairs of associations; $t_a * t_b$; $t_a * t_c$; $t_b * t_c$, etc. The fact that t_a, t_b and t_c belong to $P_a X P_b X P_c$ does not imply that $t_a * t_b * t_c$.

The same information is given by a simple punched card system. We can design a card for each term or "page" with 5000 dedicated positions for $t_1, t_2, \ldots, t_{5,000}$, numbered from 1 to 5000. The seventh hole, for example, is punched on every card representing a P to which t_7 belongs, including, of course, P_7. Thus any P can be represented by a card on which all the members of P are punched in fixed pre-assigned positions.

The set of associations

$$(t_a, t_b, t_c) * (t_a, t_b, t_c, \ldots)$$

will be given immediately by the punched holes common to all three cards P_a, P_b, and P_c since each common hole represents a member of $P_a \times P_b \times P_c$. But suppose we wished to use this same simple method of superimposition to find $t_a * t_b * t_c$. We would then have to construct P_1, P_2, etc. by listing on any P_k all the 2-term combinations rather than the single terms associated with t_k. On any single "pages" the number of actual 2-termed combinations $(t_1 * t_2)$, $(t_2 * t_3)$, $(t_1 * t_3)$, etc. might not be appreciably larger than the number of single terms, but in order to use superimposition of punched cards to find any associated pair associated with a third term, $(t_a * t_b) * t_c$, we would require a dedicated position on each card, not for 5,000 terms but for each of the $\frac{5000 \times 4999}{2}$ possible or potential pairs. A card with that many dedicated positions is a practical absurdity.

We certainly will find it necessary to go beyond such chains and to discover whether in any specific case the associations $t_a * t_b$, $t_a * t_c$, $t_b * t_c$ are in S because t_a, t_b, t_c are contained in a single set I_1, that is, $(t_a * t_b * t_c)$ or

whether the associations represent three separate sets

$$I_1\ (t_a * t_b)$$

$$I_2\ (t_a * t_c)$$

$$I_3\ (t_b * t_c)$$

It appears, at this stage of our investigations that the indexing machine we have designed for other purposes can also be used to answer such questions quickly and automatically, and the next report will describe the indexing machine with particular reference to its use in problems involving the association of ideas.

BOUNDARY (t_a)

Aerodynamics
Air
Airfoils
Alloys
Alpha
Aluminum
Analysis
Angle
Anisotropy
Arrangements, atomic
"Atlas"
Atomic

Bends
Blades
Bodies
Boundaries

Calculation
Cascades
Cobalt
Coefficient
Compressible
Conditions
Conduction
Configurations
Continuous
Convection
Copper
Creep
Crystals

Data
Decay
Deflection
Density
Determination
Diagrams
Diffusers, supersonic
Diffusion
Distribution
Drag
Duct

Ejectors, jet
Electromagnetic
Errors
Estimation
Etches
Exchange
Experiments, Nikuradse

Factor
Feshback-Weisskopf-
Peaslee theory
Flat
Flow
Fluctuations
Fluids
Formulas
Friction

Gas
Gradient
Grain

Heat
Helium II
Hermes
High
Hydraulics

Ice
Impulsive
Incidence
Index
Induced
Infinite
Incompressible
Inhomogeneous
Integral
Interaction
Ionization
Iron
Isothermal
Isotropic

Jets

Laminar
Layers
Lift
Liquid
Low

Measurements
Mechanics
Medium
Metals
Method
Molybdenum
Momentum
Motion
Movement

Nickel
Nikuradse experiments
Nozzles

Optics
Optimum

Parabolic
Parachutes
Peaslee theory
Penetration
Perturbation
Phases
Photographs
Pipes
Pitot
Plane
Plates
Point
Poiseville flow
Precipitation
Prediction
Pressure
Prevention
Propagation

Q/VF technique

Radiation
Radioactive
Recovery
Reflection
Refractive
Regions
Revolution

Schlieren
Secondary
Sections
Secular ice
Self-induced
Separation
Ships
Shock
Silicon
Silver
Skewed jets
Skin
Slamming
Smoke
Smooth
Snell's law
Soil
Solid
Solutions
Speeds
Stability
Stall
Stress; stressing
Subsonic
Suction
Supersonic
Surface
Systems

Tandem
Tank
Technique, Q/VF
Temperature
Ternary
Tests
Theory
Thermodynamics
Tracer
Transfer
Transformations
Transition
Troposphere
Tubes
Tunnels
Turbomachines
Turbulence
Turning

USSR

Velocity
Viscous
Visual
Visualization
Volume

Wake
Water
Waves
Weak
Weisskopff theory
Wind

X-ray

Zinc

EXHIBIT 1

WAVES (t_b)

Acoustics
Aerodynamics
Air
Airborne
Aircraft
Airfoils
Analysis
Antennas
Antimony
Apparatus
"Atlas"
Attenuation

Bacteria
Basin
Beams
Bending
Blocking
Bodies
Boundaries

Cadmium
Calculations
Cantilever beam
Cavities
Cerenkov radiation
Charts
Circular
Climatology
Clouds
Coating
Cockpit
Combustion
Computer; Computing
Compressible
Conducting
Convection
"Crest" (ship)
Crystals
Curves

Damped
Data
Debye
Delta
Design
Detonation
Dielectric
Diffusers, supersonic
Digital
Discontinuities
Dispersion
Disturbance
Domes
Drag

Earthquakes
Ejectors, jet
Elastic; Elasticity
Electromagnetic
Electrostatic
Elevator
Equations
Estimation
Exchange
Exciters

Fabrication
Field
Filled, Dielectric
Fine
Flat
Flow
Fluids
Flush-mounted
Fourier integral
Frequency

Generators
Glow
Great Britain
Grinding

Heat
Helium
High
Hobbing, hot
"Horizon" (Ship)
Horizontal
Humidity
Hydraulics
Hydrogen
Hyperfine structure

Incompressible flow
Inhomogeneous
Integral
Intensity
Interaction
Interdigital
Interference
Ionization
Isotropic

Jets

Kamchatka peninsula

Laminar
Layers
Lethality
Limits
Linear
Lines
Load
Long
Low

Magnetrons
Maneuvering
Mantle
Mathematics
Mechanics
Medium

Meteorological
Method
Micrococcus pyogenes
(var. aurens)
Microwaves
Millimeter
Mixing, jet
Model
Modes
Moisture
Molecular
Momentum
Motion
Mounted

Nonlifting
Nozzles
Nuclear

Observations
Oceans
"Ordvac"
Oscillators

Pacific Ocean
Palisades seismograph
Parachutes
Pattern
Photography
Piezoelectric
Plane
Plates
Poiseville flow
Polar
Pond
Pressure
Propagation

Q/VF technique
Quadrupole lines

Radiation
Radio
Radiosondes
Ratio
Rayleigh
Reflection
Reflex
Oscillator
Research
Resonators
Revolution

Schlieren
Seismographs
Shell
Shielding
Ships
Shock
Sine wave
Single
Sirens
Skewed jets
Slamming
Slope
Slots
Soil
Spectrometers
Spherical
Strength
Structure
Stress; Stressing
Subsonic
Superrefraction
Supersonic
Surface
Suspension
Synoptic charts

Table
Tail
Tapered
Technique, Q/VF
Temperature
Tests
Thickness
Theory
Timoshenko's
equations
Transfer
Transmission
Travelling waves
Troposphere
Tubes
Tunnels
Turbulence

Ultra-High Frequency
Unsteady
Upper
Utilization

Varying
Vibrations
Visualization

Wake
Water
Waveguides
Waves
Weak
Westerlies
Whistles
Wind
Wing

Yacht

Zinc

EXHIBIT 2

LAMINATION; LAMINAR (t_c)

Acetylene
Adhesives
Air
Analysis
Annealing
Araldite
Axial

Benzyl alcohol
Bonding
Boundaries
Breakdown, dielectric
Burners
Burning

Calibration
Carbon
Channels
Cloth
Coatings
Compressible
Constant
Convection
Crazing
Curing

Deterioration
Development
Dielectric
Displacement
Distillation
Dowanols

Effects
Elastomer
Electrical
Epon
Erosion
Errors
Estimation
Ethylurea

Filters
Flames
Flat
Flow
Flowmeters
Fluids
Friction

Gases
Glass

Heat
High
Hydrogen
Hypersonic

Improvement
Insulating
Integral
Interaction
Interlayers

Jets

Layers
Loads

Machine
Materials
Mathematics
Meter
Methacrylate
Method
Mica
Mixing
Moisture
Molding
Momentum
Monoxide
Motor

Natural
Neoprene
Nylon

Oxygen

Parallel
Phenol
Pitot
Plastic
Plasticizers
Plates
Plexiglas
Point
Polyethylene
Polyesters
Polytetrafluorethylene
Polyvinylbutyral
Porous
Pressure

Rain
Reflection
Regions
Resins
Resistance
Rocket

Separation
Shock
"Silastic 240"
Silicones
Skin
Speeds
Stability
Streams
Strength
Stress; stressing
Subsonic
Supersonic

Surface
Symmetric

Temperature
Tests
Thickness
Transfer
Transformations (Math)
Transition
Treatment
Tubes
Tunnel
Turbulence

Urea

Velocities

Walls
Waves
Weak
Wettability
Wind

EXHIBIT 3

BOUNDARY (t_a) and WAVES (t_b)

Aerodynamics
Air
Airfoils
Analysis
"Atlas"

Bodies
Boundaries

Calculations
Convection
Crystals

Data
Diffusers, supersonic
Drag

Ejectors, jet
Electromagnetic
Estimations
Exchange

Flat
Flow
Fluids

Heat
High
Hydraulics

Incompressible flow
Inhomogeneous
Integral
Interaction
Ionization
Isotropic

Jets

Lamination; laminar
Layers
Low

Mechanics
Medium
Method
Momentum
Motion

Parachutes
Plates
Poiseville flow
Pressure
Propagation

Q/VF technique

Radiation
Reflection
Revolution

Schlieren
Shock
Skewed jets
Soil
Stress; stressing
Subsonic
Supersonic
Surface

Technique, Q/VF
Temperature
Tests
Theory
Transfer
Troposphere
Tubes
Tunnels
Turbulence

Visualization

Wake
Water
Weak
Wind

Zinc

EXHIBIT 4

BOUNDARY (t_a) and WAVES (t_b) and LAMINATION; LAMINAR (t_c)

Air
Analysis

Boundaries

Compressible
Convection

Estimation

Flat
Flow
Fluids

Heat
High

Integral
Interaction

Jets

Lamination; laminar
Layers

Method
Momentum

Plates
Pressure

Reflection

Shock
Stress; stressing
Subsonic
Supersonic
Surface

Temperature
Tests
Transfer
Tubes
Tunnels
Turbulence

Waves
Weak
Wind